EPCC's Exascale journey: a retrospective of the past 10 years and a vision of the future

Department: Head
Editor: Name, xxxx@email

EPCC's Exascale journey: a retrospective of the past 10 years and a vision of the future

M. Weiland
EPCC, The University of Edinburgh

M. Parsons
EPCC, The University of Edinburgh

***Abstract*—The preparations for Exascale computing started over a decade ago and EPCC, the supercomputing centre at the University of Edinburgh, was at the heart of these efforts from the beginning. In this article, we revisit some of the key research projects that have paved our Exascale path for the past 10 years and look to the future.**

EXASCALE supercomputing is becoming a reality worldwide. The next few years will see a series of announcements from the world's major economies as they unveil a succession of supercomputers capable of delivering an Exaflop of numerical performance. While being able to deliver the capability of Exascale computing is a somewhat arbitrary challenge, it has proven a difficult challenge to meet and the real work began in the first decade of the new millennium.

EPCC[1] is the UK's largest supercomputing centre with many international links. Established in 1990 it has grown to be a large centre with interests spanning supercomputing, AI and data science. In 2021 it has 120 staff and hosts many of the UK's national supercomputer services including the 27 Petaflop ARCHER2 system. In 2008, when this story begins, EPCC was hosting the 208 Teraflop HECToR national HPC service and beginning to prepare for its first Petaflop system.

In June of that year, the Top500 announced that the Roadrunner system at Los Alamos National Laboratory in the USA had broken the Petaflop barrier for the first time. In parallel, an activity was already underway to think about the next frontier - the Exascale - with a key report by the DARPA Exascale Study Group into *Technology Challenges in Achieving Exascale Systems* [1], being published later that year. This DARPA report, and initial activities in Japan focused on the "Post-K Project"[2] which began even before the completion of the K computer, led to increased interest in Europe and the de-

[1]https://www.epcc.ed.ac.uk

[2]This is the system we now know as the Fugaku supercomputer at Riken CCS in Kobe

velopment of the European Exascale Software Initiative (EESI) [2], active from 2012 to 2015. EESI was aligned with the International Exascale Software Project (IESP) [3] in the USA, which was primarily active from 2009 to 2012.

By 2010 the European Commission were becoming increasingly interested in the Exascale challenge and opened the first call for proposals that year which focussed on *Exascale computing, software and simulation*. Three large-scale projects were funded in 2011 and EPCC led one of them - the CRESTA project which is described below. CRESTA, Mont-Blanc and DEEP set the scene for a series of projects that EPCC has led or been involved in over the past decade. During this period EPCC has been involved in many of the key Exascale system development challenges outlined in the DARPA report, namely: Energy and Power, Memory and Storage, Programming Models, and Exascale Algorithms and Software. It is this set of key projects, outlined below, that have formed EPCC's Exascale journey - a journey that we hope will lead to the opening of the UK's first Exascale Supercomputer service in 2024.

Exascale algorithms and software

From the start of the Exascale development story the concept of co-design was discussed in great detail. EPCC had been involved in the hardware co-design of the QCDOC computer in the 1990s with IBM, Columbia University and Brookhaven National Laboratory in the USA.

The opportunity to co-design hardware is rare, and with Exascale hardware a long way off in 2011, EPCC decided to focus on the concept of software co-design. Rather than co-designing hardware and software applications, we made assumptions about the likely level of parallelism, networking and processor performance of proposed Exascale systems and considered how software applications and "systemware" (meaning operating systems, communication protocols, compilers, debugging tools and software libraries) could be designed to work on them. This was the focus of the CRESTA[3] project - the first true Exascale project that EPCC led from 2011 - 2014.

CRESTA focused on six applications from the domains of science and engineering (Elmfire, GROMACS, HemeLB, ECMWF IFS, Nek5000 and OpenFOAM) and considered how, or indeed if, they could become Exascale applications. A key part of CRESTA was its focus on the concepts of *incremental change* and *disruptive change*.

The past history of supercomputing as we moved into the Terascale and Petascale periods predominately focused on making incremental changes to applications so that they could take advantage of new processor technologies and scale to higher levels of parallelism. With Exascale designs at that time suggesting levels of parallelism of 100 to 1,000 times that of Petascale systems, it was clear many applications would require to be completely rewritten - a process that would be highly disruptive and costly.

During this period EPCC also increased its international collaboration activities. Building relationships in the USA, Japan and Russia along with strengthening collaboration with all of the major European supercomputing centres. A result of this was the creation in 2013 of the EASC (Exascale Applications and Software) series of conferences in partnership with the KTH Royal Institue of Technology in Stockholm.

Power and energy efficiency

As scientific simulations and the hardware that they run on, move to a bigger and bigger scale, the amount of power and energy that is consumed by these simulations is increasingly becoming an area of interest (and potentially concern). EPCC secured European funding to lead a project called ADEPT[4], which was active from 2013 to 2016. The project had a particular focus on measuring and understanding power consumption in parallel software and hardware. Being able to accurately measure the power draw of the individual components that make up a compute node is key to understanding where and how energy is used, and this was the driver behind the development of a fast (1 million samples per second) power measurement infrastructure. Figure 1 shows the type of information that was derived from this high-resolution measurement infrastructure: the iterative pattern of the stencil computation is clearly visible in the change of the power that is drawn by the CPU. For fast iterations, frequent

[3]https://cordis.europa.eu/project/id/287703

[4]https://cordis.europa.eu/project/id/610490

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2021.3119101, Computing in Science & Engineering

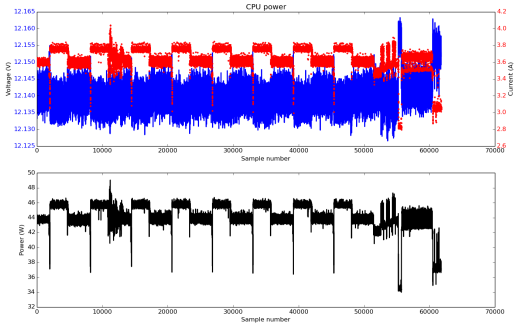

Figure 1. CPU power draw variability as measured by *ADEPT*, showing the saw-tooth power draw pattern of a simple iterative stencil operation.

measurements are important - sampling at too low a frequency means that changes can easily be missed.

The ADEPT project used the power measurement infrastructure principally on small, single node, systems [5], but the knowledge and understanding gained from the research at the small scale was later used to investigate the trade-offs between performance and energy use on larger scale systems as well [4]. The aim was to try and quantify the impact that our choices of programming models, parallel algorithms, compilers and hardware have on the efficiency (where "efficiency" relates to power, energy and runtime in equal measure) of the applications we run, and to raise awareness that these choices do indeed have an impact beyond performance. Energy usage is directly proportional to runtime, and thus energy efficiency also relies on good parallel efficiency. A system that has high power consumption but also delivers high parallel efficiency and performance should be energy efficient. The idle power consumption of a system was found to be a crucial (and often dominant) factor in any optimisation efforts;r this is an area where many advances have been made in recent years with idle components moving to low (or no) power.

Interoperability of programming models

In order to deal with the greater levels of parallelism and the deep memory hierarchies exposed in today's (and future) supercomputers, application developers are increasingly moving away from pure message-passing (MPI) and towards mixed-node parallelization techniques. MPI remains the principal parallel programming model for distributed memory, however on-node and on-accelerator models such as OpenMP are becoming a key part of preparing applications for the Exascale as they enable a finer-grained expression of parallelism which can be crucial to extract performance from the hardware. The INTERTWinE project[5], which was active between 2015 and 2018, studied the challenges of interoperability between different programming models, both widely used models such as MPI and OpenMP, as well as (at the time) lesser known task-based and partitioned global address space (PGAS) models. Based on its research and findings, the project developed a number of Best Practice guides, with tips and tricks for how to combine different programming models, how to migrate legacy code to use new parallel programming techniques, and what pitfalls to avoid.

I/O performance

Another important performance aspect of large-scale scientific applications is their ability to read and write large volumes of data. Writing data in parallel can quickly become a performance limiting factor at scale, especially if writing includes metadata operations such as creating directories. In 2015, EPCC embarked on a four-year project, called NEXTGenIO[6], to explore the use and usefulness of a new technology developed by Intel and Micron, Optane Data Centre Persistent Memory modules (DCPMM). The aim of the project was to assess how this technology might benefit supercomputing applications that struggle with I/O performance, or indeed applications that have very high memory demands, because DCPMM can act both as very fast storage or slower persistent memory, depending on its mode of operation. To that end, the NEXTGenIO project developed a 34-node prototype system with 3TB of DCPMM per node, as well as system software to evaluate the hardware for a range of applications, from traditional numerical simulations to machine learning applications. We tested applications at a range of scales and with a variety of different system configurations,

[5]http://www.intertwine-project.eu
[6]http://www.nextgenio.eu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2021.3119101, Computing in Science & Engineering

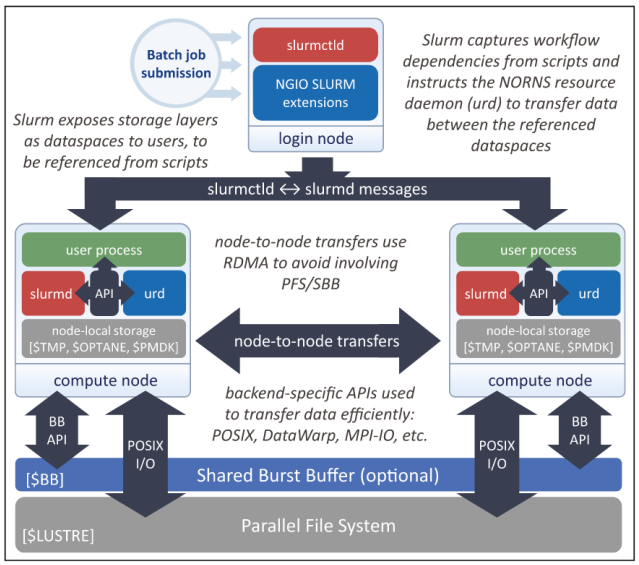

Figure 2. Example of system software research undertaken as part of the NEXTGenIO project, in this case focusing on supporting data-driven workflows (picture from [7]).

and found the I/O performance to be extremely impressive, achieving more than 70GB/s write performance on 16 nodes as part of a weather simulation [6].

Figure 2 shows the project's research around system software support. The research shown here was jointly undertaken by EPCC and Barcelona Supercomputing Centre, and investigated how workflows with data dependencies can be supported on a system with persistent memory [7]. The SLURM resource manager was modified to be aware of the persistent memory, and it was extended to enable users to reconfigure and reboot compute nodes with specific configurations for their jobs.

The NEXTGenIO prototype system, which was built by Fujitsu Germany as part of the project, is hosted at EPCC's data centre.

Digital Twins

One of the drivers behind Exascale is the ability to create "Digital Twins": computational simulations that can replicate the whole-system behaviour of (physical) processes or entities with high fidelity. They are virtual prediction and modelling tools, built from computational simulation and machine learning (ML) applications, and they have the potential to enable massive reductions in the cost and time required to develop new products, ensure safe operation of large scale systems, undertake timely predictive maintenance, and understand real world impacts of systems and events. From a computational point of view there are core challenges associated with large-scale Digital Twins that need to be overcome: high performance, excellent scalability, robustness and accuracy in the face of unprecedented complexity are common requirements.

The Strategic Prosperity Partnership[7] "Advanced Simulation and Modelling of Virtual Systems" (ASiMoV) is a 5-year project that aims to create a Digital Twin of a complete gas-turbine (i.e. an aircraft engine) during operation. ASiMoV, a multi-partner collaboration, started in late 2018 and is jointly led by EPCC and Rolls-Royce. The ultimate aim of the project is to move the quality of the simulation towards a fidelity and accuracy that is sufficient to be able to perform virtual certification of an aircraft engine, i.e. certifying engines entirely based on computation. This involves simulating coupled models of the fluid dynamics, combustion, structural and thermo-mechanics, a huge challenge for all aspects of computing. ASiMoV integrates the multi-faceted challenges of large-scale industrial

[7]Prosperity Partnerships are business-led research partnerships between leading UK based businesses and their long-term strategic university partners

simulations with fundamental research into areas such as extreme scaling, physical modelling, cost of computing, security of data and trust in the simulation results.

Underpinning technologies

Many of the use cases presented in the projects above have one common underpinning aspect: their reliance on computational meshes to discretize their problem domains. Unfortunately many operations that involve meshes (such as mesh generation, mesh adaptivity, mesh partitioning) do not scale to the same levels as the computation performed on the meshes. The ExCALIBUR working group ELEMENT[8] is a short project that aims to distill the challenges around "meshing" *for* and *at* the Exascale in a Vision Paper, and it will propose a strategic research agenda that will drive meshing closer to the level that is required at the Exascale.

The meshes required for Exascale simulations, will model problems with extreme geometric complexity and levels of refinement, will be very large (on the order of 10^9 cells and above), and contain cells that may differ in size by many orders of magnitude to be able to faithfully resolve any underlying physics at their appropriate scales. Meshing and geometry management remain a significant performance bottleneck and they pose a challenging obstacle that must be overcome to enable Exascale simulations. Another challenge that must be addressed is that of mesh quality, which is often insufficient for high-fidelity simulations.

The ELEMENT Vision Paper will be published in late 2021.

Novel hardware

Getting ready for the Exascale also involves experimenting with novel hardware and working on understanding if (and how) this hardware might play a role in extreme-scale scientific computing. The ExCALIBUR project recently funded a number of hardware testbeds through its "Hardware and Enabling Software"[9] programme. EPCC hosts two systems that have been supported by the programme: a Cerebras CS-1 Wafer Scale Engine; and a field-programmable gate array (FPGA) testbed.

The Cerebras CS-1 is the first such system in Europe, and it enables performance and usability exploration for UK academic and industrial users. The majority of the system was funded by the University of Edinburgh as part of the Edinburgh International Data Facility which EPCC hosts, but support from ExCALIBUR allows for a more general access service to be provided to researchers from across the wider computational science and AI community in the UK. The Wafer Scale Engine is the world's largest processor, at over $46,000$ square millimetres, with 1.2 trillion transistors, $400,000$ processor cores, 18 gigabytes of SRAM, and an interconnect between processors capable of moving 100 million billion bits per second. The CS-1 system is focused on neural network training and the hardware is integrated with common machine learning frameworks such as TensorFlow and PyTorch2, providing the potential for extreme performance for a wide range of machine learning training tasks. The CS-1 became operational in May 2021.

The second testbed is aimed at researchers wanting to explore FPGA technology for their scientific and data-science applications, and investigate its potential performance and power advantages. The testbed will present a unique resource within UK academic computing: it will provide access to the Versal Adaptive Compute Acceleration Platform (ACAP) technology from Xilinx, which includes their revolutionary AI engines. It will also include technologies such as high bandwidth (HBM2) and non-volatile (NVRAM) memory, as well as multiple networking options, including a high performance node-level network and direct FPGA-to-FPGA networking to enable direct comparison and assessment of the relative merits of both approaches. The system will contain multiple families of FPGA, allowing the evaluation of a range of technologies. The FPGA testbed will come online in Autumn 2021.

CONCLUSION

EPCC, in common with many supercomputing centres world-wide, is now over 13 years into its preparations for the Exascale. As we have

[8]https://epcced.github.io/ELEMENT/
[9]https://excalibur.ac.uk/

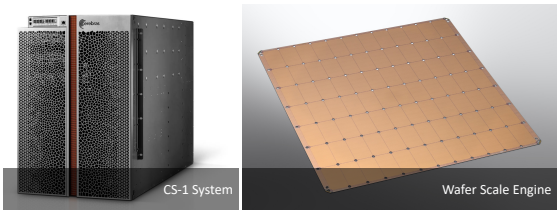

Figure 3. The Cerebras CS-1 system and its Wafer Scale Engine. (*Picture courtesy of Cerebras*)

outlined in this article, over the past decade we have been involved in a wide variety of projects focused on many of the key problems the Exascale poses. We have only summarised some of the key projects here. Our projects have grappled with key hardware and software challenges and, we hope, have supported the combined international efforts to get to the Exascale early in this decade.

As we look to the Exascale future, EPCC has been focusing since 2018 on building the necessary hosting environment for an Exascale system. In 2020 we completed a new state-of-the-art computer room at our data centre in Edinburgh. We have also invested in a new 30MW power supply to our site (increasing our site power from 8MW to 38MW). By 2022 we will be ready to host the UK's first Exascale supercomputer and hope to make this a reality for the UK's science agency, UK Research & Innovation, by 2024.

The Exascale represents a major milestone in the history of supercomputing. We are certain that the scientific and industrial research and innovation that forthcoming systems enable will be truly transformational and touch all of our lives.

ACKNOWLEDGMENT

The projects presented in this article received funding from various programmes: the European Commission's Seventh Framework Programme, Grant Agreements no. 287703 (CRESTA) and 610490 (ADEPT); the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreements no. 671602 (INTERTWinE) and 671951 (NEXTGenIO); the UK Engineering and Physical Sciences Research Council, Grant IDs EP/S516107/1 (ASiMoV) and EP/V001345/1 (ELEMENT).

REFERENCES

1. K. Bergman, S. Borkar, D. Campbell et al., "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems" P. Kogge, Editor & Study Lead, DARPA, 2008.
2. European Exascale Software Initiative [Online]. Available: http://www.eesi-project.eu (URL)
3. International Exascale Project [Online]. Available: https://www.exascale.org/iesp (URL)
4. M. Barefore, N, Johnson and M. Weiland, "On the Trade-offs between Energy to Solution and Runtime for Real-World CFD Test-Cases", Exascale Applications and Software Conference 2016 Proceedings. ACM, 2016. doi: 10.1145/2938615.2938619
5. M. Weiland and N. Johnson, "Benchmarking for power consumption monitoring", *Computer Science - Research and Development*, Vol. 30, No. 2, 01.05.2015, p. 155-163. doi: 10.1007/s00450-014-0260-1
6. M. Weiland, H. Brunst, T. Quintino, et al., "An Early Evaluation of Intel's Optane DC Persistent Memory Module and its Impact on High-Performance Scientific Applications" *SC'19 Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM, 2019. p. 1-19. doi: 10.1145/3295500.3356159
7. A. Miranda, A. Jackson, T. Tocci, et al., "NORNS: Extending Slurm to Support Data-Driven Workflows through Asynchronous Data Staging", *2019 IEEE International Conference on Cluster Computing (CLUSTER)*, Albuquerque, New Mexico, USA (23-26 September 2019). doi: 10.1109/CLUSTER.2019.8891014

Dr Michèle Weiland is a Senior Research Fellow at EPCC. She specialises in novel technologies for extreme scale parallel computing, leading EPCC's technical work in the ASiMoV Strategic Prosperity Partnership with Rolls-Royce. She is a Co-I of the ExCALIBUR Design and Development Working Group (DDWG) project ELEMENT (EP/V001345/1), which addresses the use case of meshing for and at the Exascale, and of the Cirrus II Tier 2 HPC service (EP/T02206X/1). She also leads on EPCC's involvement in the Catalyst UK programme, a partnership with HPE and Arm to accelerate the adoption of the Arm ecosystem. She is the EPCC PI on a number of research grants, including the EC Horizon 2020 projects "HPC in Wind Energy" (grant ID 828799) and SAGE2 (grant ID 800999). Michèle is a member of the EPSRC e-Infrastructure Strategic Advisory Team and an Associate Director of the Arm HPC User Group. Contact her at m.weiland@epcc.ed.ac.uk.

Professor Mark Parsons is the Director of EPCC. Following a PhD in Particle Physics at CERN on the LEP accelerator he joined EPCC as a software programmer in 1994. He became EPCC's Director in 2016. Since 2020 he has worked part-time as Director of Research Computing for the UK Government's Engineering & Physical Sciences Research Council where he is leading the UK Exascale Supercomputer project. He has wide interests in supercomputing and is well known for his work with industry, particularly small-to-medium sized companies, through his leadership of the EC-funded Fortissimo projects (grant IDs 609029 and 680481). He has a led a number of EPCC's Exascale projects including the EC Framework 7 project "CRESTA" (grant ID 287703) and the EC Horizon 2020 project "NEXTGenIO" (grant ID 671591). His is the 2021 Chair of the ACM Gordon Bell Prize Committee. Contact him at m.parsons@epcc.ed.ac.uk.